AD________________

Award Number: W81XWH-10-1-0551

TITLE: The Role of Sox4 in Prostate Cancer Metastases

PRINCIPAL INVESTIGATOR: John Prensner, B.A.

CONTRACTING ORGANIZATION: University of Michigan
Ann Arbor, MI 48105

REPORT DATE: September 2011

TYPE OF REPORT: Annual Summary

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

1. REPORT DATE September 2011	**2. REPORT TYPE** Annual Summary	**3. DATES COVERED** 1 September 2010 – 31 August 2011

4. TITLE AND SUBTITLE The Role of Sox4 in Prostate Cancer Metastases	**5a. CONTRACT NUMBER**
	5b. GRANT NUMBER W81XWH-10-1-0551
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S) John Prensner	**5d. PROJECT NUMBER**
	5e. TASK NUMBER
E-Mail: diannal@umich.edu	**5f. WORK UNIT NUMBER**
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) University of Michigan Ann Arbor, MI 48105	**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) U.S. Army Medical Research and Materiel Command Fort Detrick, Maryland 21702-5012	**10. SPONSOR/MONITOR'S ACRONYM(S)**
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

High-throughput sequencing of polyA+ RNA (RNA-Seq) in human cancer shows remarkable potential to identify both novel markers of disease. We employed RNA-Seq on a cohort of 102 prostate tissues and cells lines and determined whether dysregulation of SOX4, a transcription factor, associates with disease. We found that SOX4 is upregulated in prostate cancer and associated with genes characteristic of embryonic stem cell gene signatures. To probe these gene signatures for extensively for new RNAs associated with disease, we performed ab initio transcriptome assembly to discover unannotated ncRNAs. We nominated 121 such Prostate Cancer Associated Transcripts (PCATs) with cancer-specific expression patterns. Among these, we characterized PCAT-1 as a novel prostate-specific regulator of cell proliferation and target of the Polycomb Repressive Complex 2 (PRC2). We further found that high PCAT-1 and PRC2 expression stratified patient tissues into molecular subtypes distinguished by expression signatures of PCAT-1-repressed target genes. Taken together, the findings presented herein identify PCAT-1 as a novel transcriptional repressor implicated in subset of prostate cancer patients.

15. SUBJECT TERMS
Prostate cancer, genomics, next-generation sequencing, gene expression, non-coding RNA

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** USAMRMC
a. REPORT U	**b. ABSTRACT** U	**c. THIS PAGE** U	UU	21	**19b. TELEPHONE NUMBER** *(include area code)*

INTRODUCTION

In the 2010-2011 funding period, I have been in the research phase of the M.D., Ph.D. Medical Scientist Training Program at the University of Michigan. My research focuses on the molecular basis of prostate cancer and especially emphasizes the transformative role of novel technologies in understanding this disease. The goal of my research is to elucidate the molecular mechanisms underlying prostate cancer and to translate these findings into novel molecular diagnostics or therapies for prostate cancer. The scope of the grant applies to the use of innovative technologies and techniques to define new molecular aspects of disease, with a particular focus on the molecular basis of aggressive, lethal prostate cancer.

BODY

Training program

Over the past year, the Department of Defense has supported my research efforts as a graduate student in the research phase of the M.D., Ph.D. Medical Scientist Training Program at the University of Michigan. My training program as a student in the Department of Pathology includes weekly seminar series that feature presentations by students as well as faculty and distinguished guest lecturers. My attendance at these seminars is required, and I have also presented my data at this forum. The Department of Pathology also has an annual research Symposium, where students present posters and attend talks by faculty and invited guests. The University of Michigan Cancer Center also holds an annual Research Symposium, at which I presented.

Mentorship

My research program is guided by my thesis mentor, Dr. Arul Chinnaiyan. Dr. Chinnaiyan is a Professor of Pathology and Urology, a Howard Hughes Medical Institute Investigator, a Doris Duke Clinical Scholar, an American Cancer Society Investigator, and a Taubman Scholar at the University of Michigan. Dr. Chinnaiyan provides a superb environment in which to learn science. He has insightful comments and keen expertise on prostate cancer research. Dr. Chinnaiyan has been instrumental in teaching me how to design and execute experiments, interpret data, and write up research reports. I meet with Dr. Chinnaiyan regularly and have frequent communication with him. I also gain guidance from the numerous other prostate cancer researchers in Dr. Chinnaiyan’s lab, including junior faculty, post-docs, and other graduate students, with whom I interact daily. Finally, my thesis committee provides regular feedback about my work in formal meetings as well as in informal settings and email communications. My thesis committee members have been supportive and highly helpful.

Conferences

I have been fortunate to attend multiple conferences during the course of the 2010-2011 funding period. I attended and gave an oral presentation at the annual DoD Prostate Cancer Research Program IMPaCT Conference in March 2011. I also attended and presented a poster at the SPORE Prostate Cancer Program Retreat in March 2011. I gave an oral presentation at the American Association for Cancer Research (AACR) Annual Meeting in April 2011. I presented a poster at the Keystone Symposium, The Changing Landscape of the Cancer Genome, in June 2011.

Mentorship experiences

Over the past year, I have worked closely to mentor undergraduate students and other trainees in the Chinnaiyan laboratory. I have specifically mentored two undergraduate students and 2 PhD rotation students during their rotations through the lab. These experiences have been extremely valuable in helping me develop my skill and comfort with mentoring other emerging scientists.

Honors/Awards

I have received several awards in the past year. My poster presentation at the University of Michigan Cancer Research Symposium received an honorable mention. My poster presentation at the SPORE Prostate Cancer Program Retreat received first prize. I was awarded the AACR-Aflac Incorporated Scholar-in-Training Award for the 2011 AACR Annual Conference.

Research Summary

The discovery of numerous non-coding RNA (ncRNA) transcripts in species from yeast to mammals has dramatically altered our understanding of cell biology, especially disease biology such as cancer. In humans, the identification of abundant long ncRNA (lncRNAs) >200 bp in length has catalyzed their characterization as critical components of cancer biology. Recently, roles for lncRNAs as drivers of tumor suppressive and oncogenic functions have appeared in prevalent cancer types, such as breast and prostate cancer.

High-throughput sequencing of polyA+ RNA (RNA-Seq) in human cancer shows remarkable potential to identify both novel disease-specific markers for clinical uses and uncharacterized aspects of tumor biology, particularly non-coding RNA (ncRNA) species. To illustrate this approach, we employed RNA-Seq on a cohort of 102 prostate tissues and cells lines. We found that aberrant expression profiles of novel tissue-specific ncRNAs distinguished benign, cancerous, and metastatic tumors, and we defined a core set of 121 novel ncRNAs whose dysregulation characterizes prostate cancer. Among these, a novel prostate-cancer specific ncRNA (termed *PCAT-1*) defined a subset of aggressive cancers with low expression of the epigenetic regulator *EZH2*, a component of the Polycomb Repressive Complex 2 (PRC2) commonly upregulated in metastatic cancers. *In vitro* chromatin immunoprecipitation, RNA immunoprecipitation, and drug treatment assays for core PRC2 genes indicated that the PRC2 complex directly binds and represses *PCAT-1*, and that *PCAT-1* transcript reciprocally binds PRC2. By contrast, *in vitro* models with high levels of endogenous *PCAT-1* transcript did not recapitulate PRC2-mediated repression, and in these cells siRNA-mediated knockdown of *PCAT-1* showed a 25 – 50% decrease in cell proliferation. Using gene expression arrays, we determined that *PCAT-1* contributes to the transcriptional regulation of genes in several key biological processes, including cell cycle. These data suggest that *PCAT-1* exhibits two biological states: a PRC2-repressed state and an active state that promotes proliferation.

Next, we showed that novel ncRNAs may serve a clinical purpose for the non-invasive detection and stratification of prostate cancer patients. We performed qPCR on patient urine samples (n=230) and found that a custom ncRNA expression signature, which includes *PCAT-1*, both diagnosed prostate cancer effectively and yielded prognostic information. Indeed, a high ncRNA expression signature value correlated with high-grade histology (Gleason score ≥7 vs. Gleason score ≤ 6; p= 0.01). Taken together, the findings presented herein establish the utility of RNA-Seq to comprehensively identify unannotated ncRNAs, such as *PCAT-1*, implicated in cancer.

Our data suggest that *PCAT-1* promotes cell proliferation, that in its inactive state *PCAT-1* is mechanistically repressed by PRC2, and that *PCAT-1* may serve as a candidate biomarker for non-invasive clinical tests. We further speculate that applying these methodologies to other diseases may reveal key aspects of disease biology and clinically important biomarkers, particularly for diseases that currently lack good non-invasive tests in fluids such as blood serum or urine.

The discovery of *PCAT-1* highlights the power of unbiased transcriptome studies to explore a rich set of lncRNAs associated with cancer. While *PCAT-1* is the first cancer lncRNA to be discovered by this method, we anticipate that many additional studies will employ this approach.

KEY RESEARCH ACCOMPLISHMENTS

- Defined the landscape of SOX4 expression across prostate cancer progression and disease subtypes (Figure 1).
- Defined SOX4 as an androgen-repressed gene (Figures 2 and 3).
- Determined global gene expression signatures associated with SOX4 and demonstrated that SOX4 knockdown results in increased E-cadherin mRNA levels (Figure 4).
- Defined novel RNA transcripts associated with SOX4 and prostate cancer (Figure 5).
- Functionally characterized novel RNA transcripts as functional molecules in prostate cancer progression (see appended manuscript, Prensner et al. *Nature Biotechnology* 2011).
- Defined one novel RNA, named PCAT-1, as a regulator of cell proliferation through transcriptional repressor of target genes. PCAT-1 is itself regulated by the Polycomb Repressive Complex 2 (see appended manuscript, Prensner et al. *Nature Biotechnology* 2011).
- Evaluated the potential for novel RNA transcripts to be utilized as novel prostate cancer diagnostics detectable in prostate cancer patient urine (see appended manuscript, Prensner et al. *Nature Biotechnology* 2011).

REPORTABLE OUTCOMES

Publications

- Cao Q, Mani RS, Ateeq B, Dhanasekaran SM, Asangani IA, **Prensner JR**, Kim JH, Brenner JC, Jing X, Cao X, Wang R, Li Y, Dahiya A, Wang L, Pandhi M, Lonigro RJ, Wu YM, Tomlins SA, Palanisamy N, Qin Z, Yu J, Maher CA, Varambally S, Chinnaiyan AM., Coordinated Regulation of Polycomb Group Complexes through microRNAs in Cancer. ***Cancer Cell*** 2011 Aug 16;20(2):187-99; PMID: 21840484
- **Prensner JR**, Iyer MK, Balbin OA, Dhanasekaran SM, Cao Q, Brenner JC, Laxman B, Asangani IA, Grasso CS, Kominsky HD, Cao X, Jing X, Wang X, Siddiqui J, Wei JT, Robinson D, Iyer HK, Palanisamy N, Maher CA, Chinnaiyan AM. Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA

implicated in disease progression. ***Nature Biotechnology*** 2011 Jul 31;29(8):742-9. doi: 10.1038/nbt.1914 PMID: 21804560

- Kim JH, Dhanasekaran SM, **Prensner JR**, Cao X, Robinson D, Kalyana-Sundaram S, Huang C, Shankar S, Jing X, Iyer M, Hu M, Sam L, Grasso C, Maher CA, Palanisamy N, Mehra R, Kominsky HD, Siddiqui J, Yu J, Qin ZS, Chinnaiyan AM. Deep sequencing reveals distinct patterns of DNA methylation in prostate cancer. ***Genome Research***. 2011 Jul;21(7):1028-41. PMID: 21724842
- Wang XS, Shankar S, Dhanasekaran SM, Ateeq B, Sasaki A, Jing X, Robinson D, Cao Q, **Prensner J**, Yocum A, Wang R, Fries D, Han B, Asangani I, Cao X, Li Y, Omenn G, Pflueger D, Gopalan A, Reuter V, Kahoud ER, Cantley L, Rubin M, Palanisamy N, Varambally S, Chinnaiyan AM. Characterization of KRAS Rearrangements in Metastatic Prostate Cancer. ***Cancer Discovery*** 2011; 1(1): OF33-41.
- **Prensner JR**, Chinnaiyan AM. Metabolism unhinged: IDH mutations in cancer. ***Nature Medicine.*** 2011 Mar;17(3):291-3

Presentations

- Keystone Symposium, The Changing Landscape of the Cancer Genome (June 2011); poster presentation
- AACR Annual Meeting (April 2011); oral presentation
- SPORE Prostate Cancer Program Retreat (March 2011); poster presentation
- Prostate Cancer Research Program (PCRP) IMPaCT conference (March 2011); poster presentation
- University of Michigan Cancer Research Symposium (November 2011); poster presentation
- University of Michigan Pathology Research Symposium(November 2011); poster presentation

Awards

- AACR-Aflac Incorporated Scholar-in-Training Award, AACR (April 2011)
- SPORE Prostate Cancer Program Retreat, First-prize poster award (March 2011)

CONCLUSION

The work funded by this project establishes the efficacy of RNA profiling to elucidate the molecular basis of prostate cancer. Through the profiling of patient tumor RNA we have no only nominated SOX4 as a prostate cancer gene, but we have also examined a co-regulated transcriptional network of novel RNA transcripts also associated with prostate cancer. We have characterized PCAT-1 as a novel noncoding RNA upregulated in prostate cancer, and we have determined that detection of ncRNAs in patient urine may be a promising avenue of non-invasive biomarkers. PCAT-1 drives cell proliferation and represses key target genes to achieve its effect. Future work would benefit from profiling non-polyadenylated RNA species as well, since these also likely have role in prostate cancer progression. In summary, this work has discovered new genes and characterized their functions in prostate cancer. This work therefore expands our knowledge and understanding of this disease, as well as nominating novel biomarkers detectable in patient urine samples.

REFERENCES

Cao Q, Mani RS, Ateeq B, Dhanasekaran SM, Asangani IA, **Prensner JR**, Kim JH, Brenner JC, Jing X, Cao X, Wang R, Li Y, Dahiya A, Wang L, Pandhi M, Lonigro RJ, Wu YM, Tomlins SA, Palanisamy N, Qin Z, Yu J, Maher CA, Varambally S, Chinnaiyan AM., Coordinated Regulation of Polycomb Group Complexes through microRNAs in Cancer. ***Cancer Cell*** 2011 Aug 16;20(2):187-99; PMID: 21840484

Prensner JR, Iyer MK, Balbin OA, Dhanasekaran SM, Cao Q, Brenner JC, Laxman B, Asangani IA, Grasso CS, Kominsky HD, Cao X, Jing X, Wang X, Siddiqui J, Wei JT, Robinson D, Iyer HK, Palanisamy N, Maher CA, Chinnaiyan AM. Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. ***Nature Biotechnology*** 2011 Jul 31;29(8):742-9. doi: 10.1038/nbt.1914 PMID: 21804560

Kim JH, Dhanasekaran SM, **Prensner JR**, Cao X, Robinson D, Kalyana-Sundaram S, Huang C, Shankar S, Jing X, Iyer M, Hu M, Sam L, Grasso C, Maher CA, Palanisamy N, Mehra R, Kominsky HD, Siddiqui J, Yu J, Qin ZS, Chinnaiyan AM. Deep sequencing reveals distinct patterns of DNA methylation in prostate cancer. ***Genome Research***. 2011 Jul;21(7):1028-41. PMID: 21724842

Wang XS, Shankar S, Dhanasekaran SM, Ateeq B, Sasaki A, Jing X, Robinson D, Cao Q, **Prensner J**, Yocum A, Wang R, Fries D, Han B, Asangani I, Cao X, Li Y, Omenn G, Pflueger D, Gopalan A, Reuter V, Kahoud ER, Cantley L, Rubin M, Palanisamy N, Varambally S, Chinnaiyan AM. Characterization of KRAS Rearrangements in Metastatic Prostate Cancer. ***Cancer Discovery*** 2011; 1(1): OF33-41.

Prensner JR, Chinnaiyan AM. Metabolism unhinged: IDH mutations in cancer. ***Nature Medicine.*** 2011 Mar;17(3):291-3

FIGURES

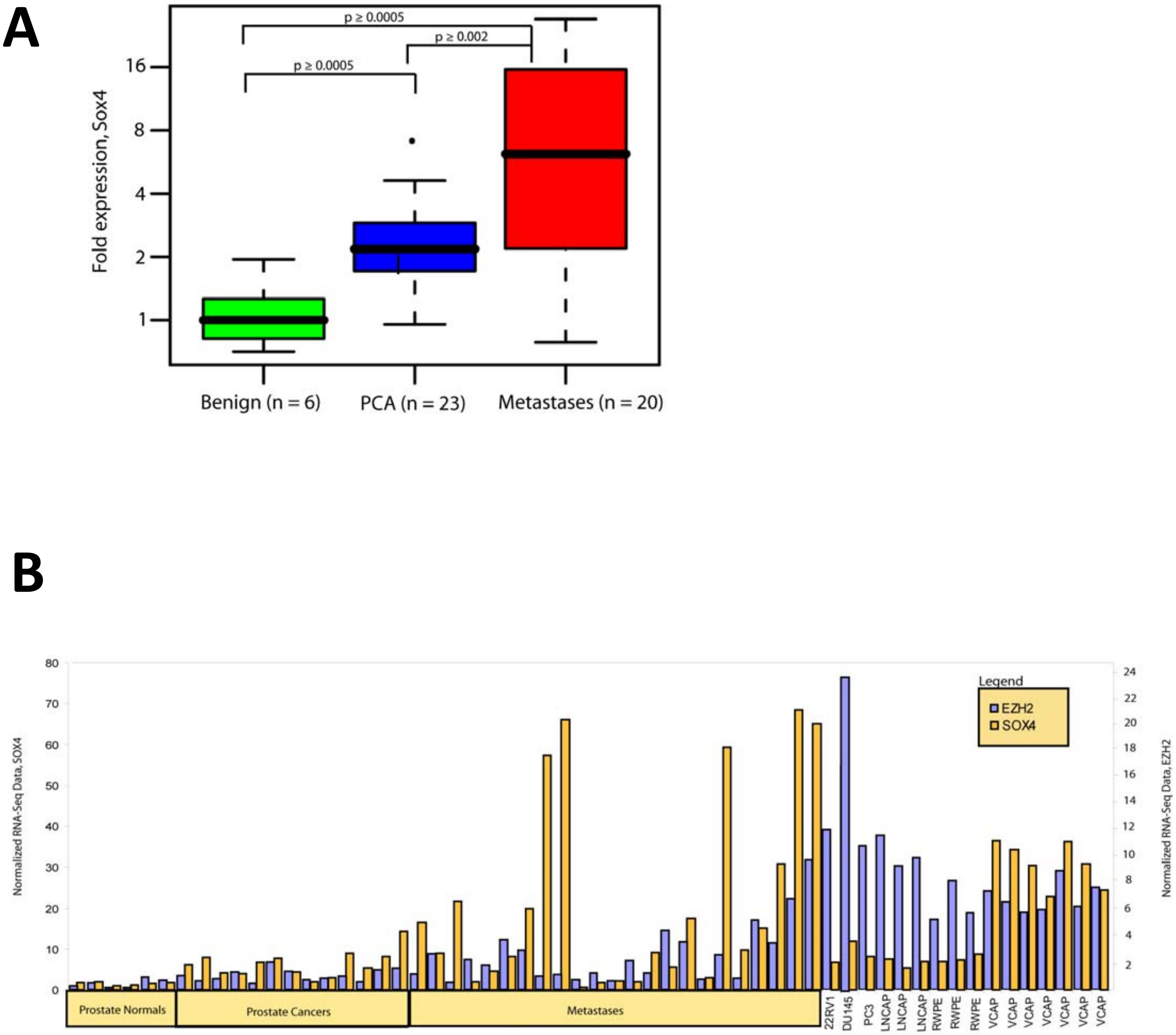


Figure 1: Nomination of SOX4 as a metastasis-associated gene. (**A**) RNA-Seq data from a cohort of prostate cancer patient samples (benign, n=6; localized cancer, n=23; metastases, n=20). SOX4 shows elevated mRNA levels in localized cancer and substantially elevated mRNA levels in metastases. (**B**) SOX4 levels in individual samples shows elevated mRNA levels in metastases. SOX4 expression is correlated with EZH2 expression, which is also elevated in metastases.

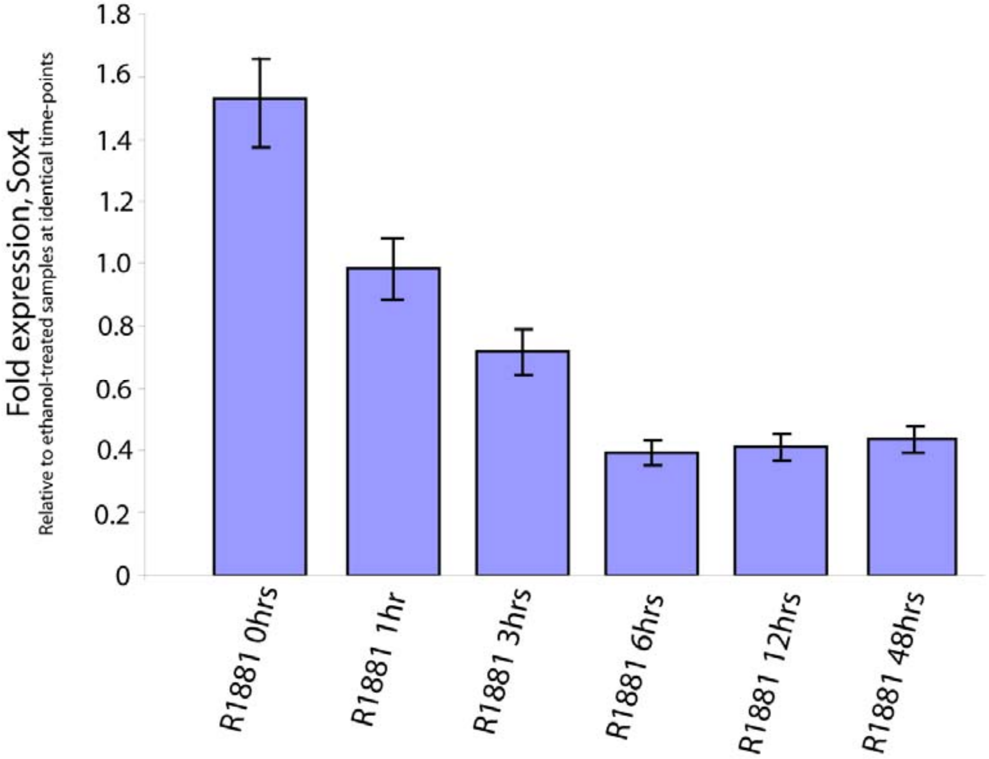


Figure 2: SOX4 is repressed by androgen stimulation. VCAP cells were starved of androgens for 48 hours and then stimulated with 10nM of R1881, a synthetic androgen. A time-course analysis of RNA expression shows decreasing SOX4 mRNA levels following induction of androgen signaling.

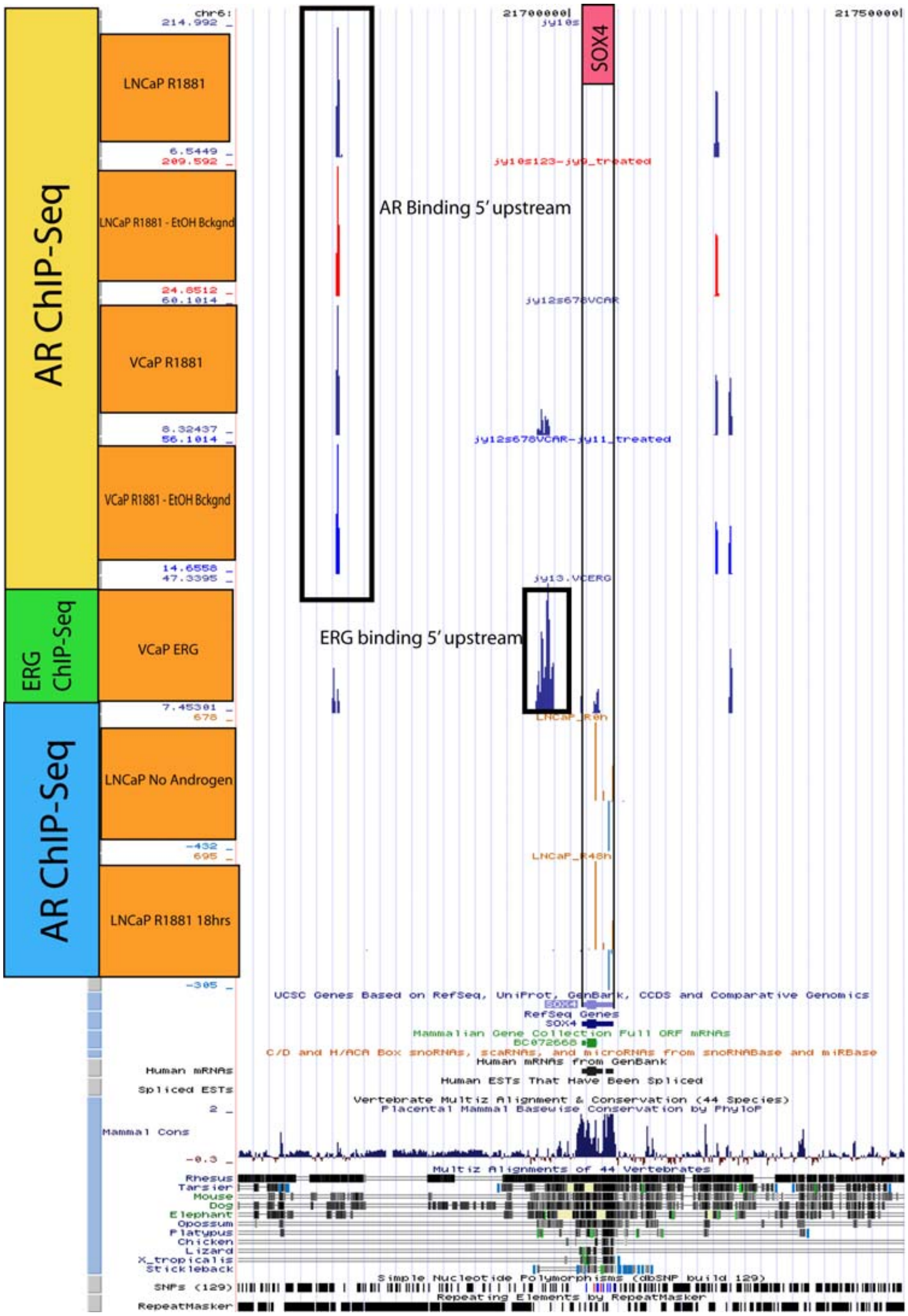


Figure 3: The SOX4 locus is bound by the androgen receptor and ERG. ChIP-Seq data for LNCaP and VCaP prostate cancer cells was performed for AR (in both cell lines) or ERG (in VCaP). AR ChIP-Seq was performed with both androgen-depleted and androgen-stimulated (R1881) conditions. AR, as well as ERG, binds the genomic locus of SOX4 upstream of the gene transcriptional start site.

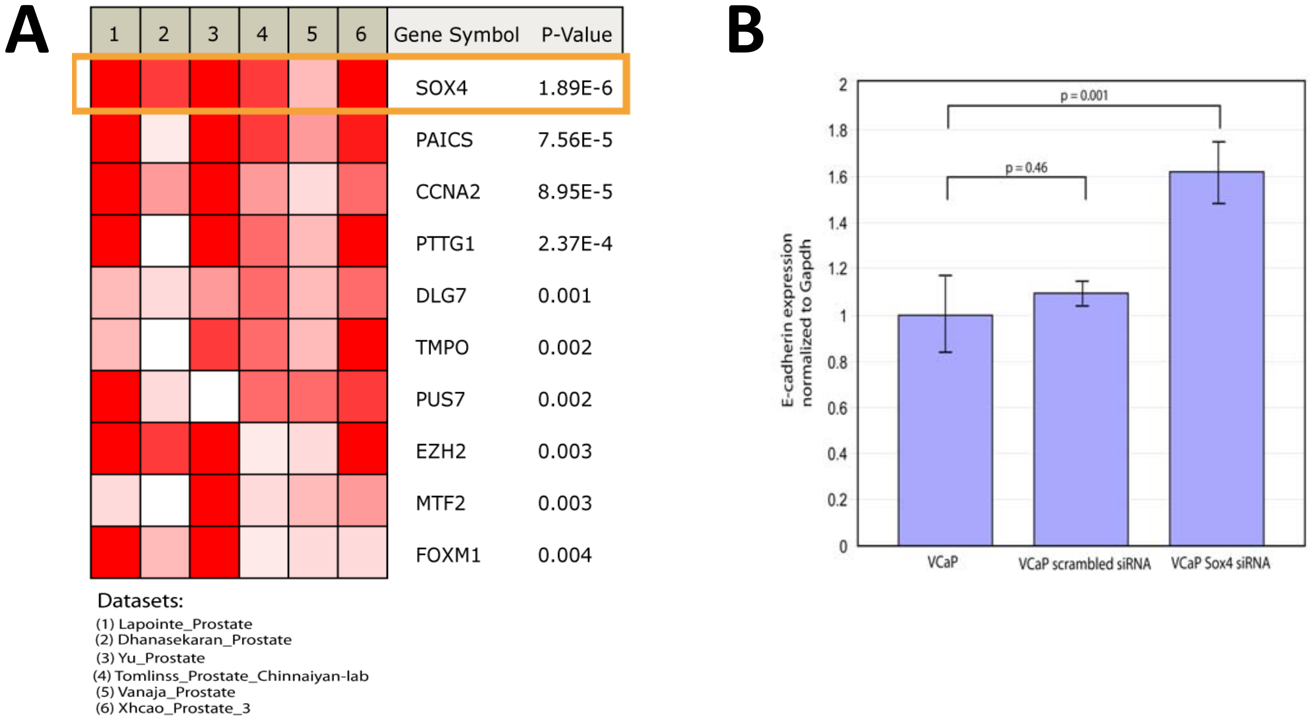


Figure 4: SOX4 is associated with metastasis gene signatures and represses E-cadherin mRNA levels. (**A**) Analysis of SOX4 expression levels using the Oncomine database nominates SOX4 as a key prostate cancer metastasis outlier gene. Analysis of SOX4 in 6 prostate cancer metastasis datasets shows that SOX4 is a commonly upregulated gene in prostate cancer metastases. (**B**) Knockdown of SOX4 in the VCaP cell line results in upregulation of E-cadherin mRNA levels.

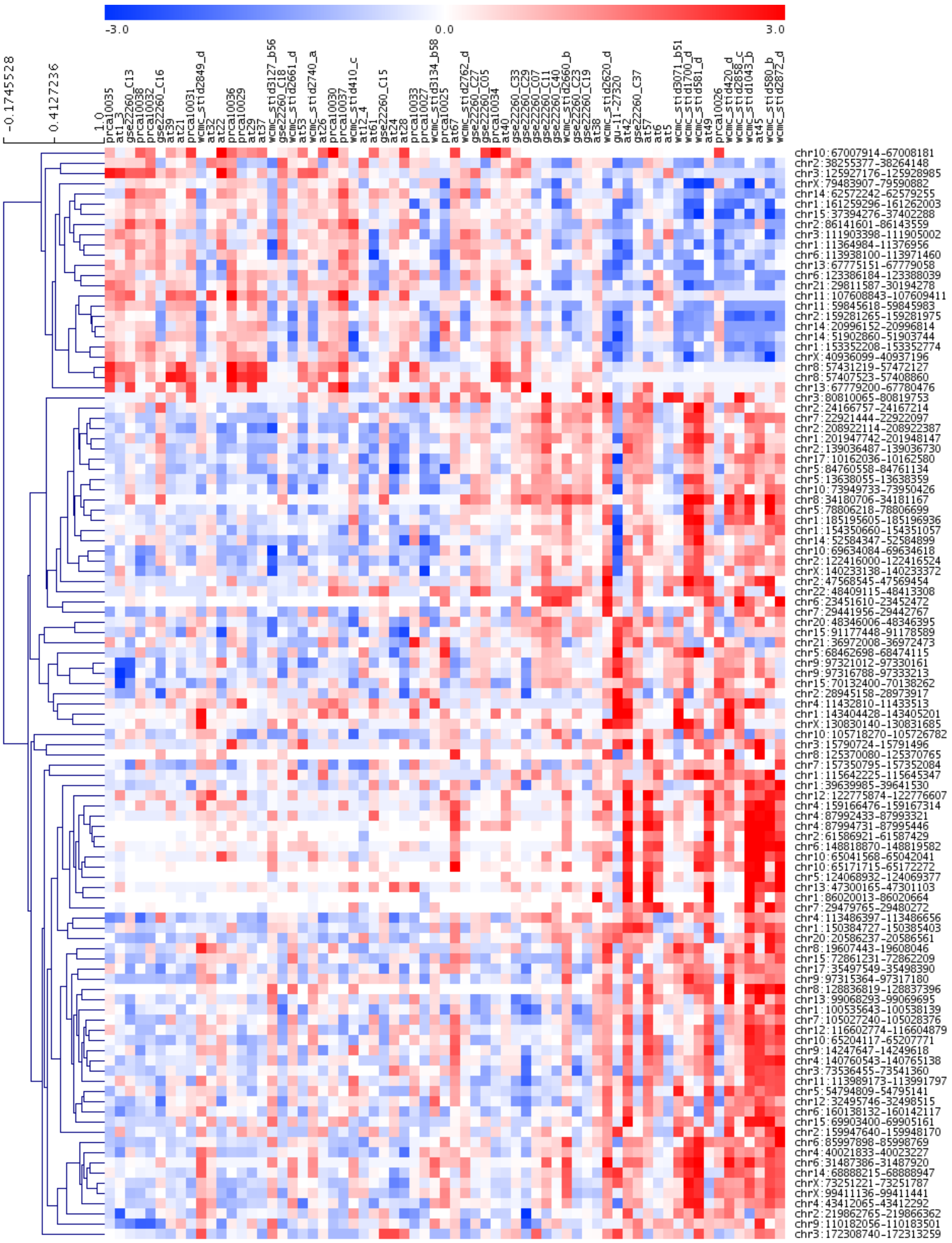


Figure 5: SOX4 is associated with novel unannotated transcripts. Prostate cancer samples analyzed by RNA-Seq were segregated according to SOX4 expression (SOX4 high vs. SOX4 low). Clustering of SOX4 with RNA-Seq predictions for unannotated lncRNA transcripts reveals a signature of novel transcripts both correlated and anti-correlated with SOX4. The samples in the heatmap above is ranked according to increasing SOX4 expression, where samples on the right side of the plot have high SOX4 expression.

nature biotechnology

Transcriptome sequencing across a prostate cancer cohort identifies *PCAT-1*, an unannotated lincRNA implicated in disease progression

John R Prensner[1,8], Matthew K Iyer[1,8], O Alejandro Balbin[1], Saravana M Dhanasekaran[1,2], Qi Cao[1], J Chad Brenner[1], Bharathi Laxman[3], Irfan A Asangani[1], Catherine S Grasso[1], Hal D Kominsky[1], Xuhong Cao[1], Xiaojun Jing[1], Xiaoju Wang[1], Javed Siddiqui[1], John T Wei[4], Daniel Robinson[1], Hari K Iyer[5], Nallasivam Palanisamy[1,2,6], Christopher A Maher[1,2] & Arul M Chinnaiyan[1,2,4,6,7]

Noncoding RNAs (ncRNAs) are emerging as key molecules in human cancer, with the potential to serve as novel markers of disease and to reveal uncharacterized aspects of tumor biology. Here we discover 121 unannotated prostate cancer–associated ncRNA transcripts (PCATs) by *ab initio* assembly of high-throughput sequencing of polyA$^+$ RNA (RNA-Seq) from a cohort of 102 prostate tissues and cells lines. We characterized one ncRNA, *PCAT-1*, as a prostate-specific regulator of cell proliferation and show that it is a target of the Polycomb Repressive Complex 2 (PRC2). We further found that patterns of *PCAT-1* and PRC2 expression stratified patient tissues into molecular subtypes distinguished by expression signatures of *PCAT-1*–repressed target genes. Taken together, our findings suggest that *PCAT-1* is a transcriptional repressor implicated in a subset of prostate cancer patients. These findings establish the utility of RNA-Seq to identify disease-associated ncRNAs that may improve the stratification of cancer subtypes.

Recently, RNA-Seq has provided a method to delineate the entire set of transcriptional aberrations in a disease, including novel transcripts not measured by conventional analyses[1–5]. To facilitate interpretation of sequence read data, existing computational methods typically process individual samples using either short read gapped alignment followed by *ab initio* reconstruction[2,3] or *de novo* assembly of read sequences followed by sequence alignment[4,5]. These methods provide a powerful framework to uncover uncharacterized RNA species, including antisense transcripts, short RNAs <250 bp or long intergenic ncRNAs (lincRNAs) >250 bp.

Although still largely unexplored, ncRNAs, particularly lincRNAs, have emerged as a new aspect of biology, with evidence suggesting that they are frequently cell-type specific, contribute important functions to numerous systems[6,7] and may interact with known cancer genes such as *EZH2* (ref. 8). Indeed, several well-described examples, such as *HOTAIR*[8,9] and *ANRIL*[10,11], indicate that ncRNAs may be essential actors in cancer biology, typically facilitating epigenetic gene repression through chromatin-modifying complexes[12,13]. Moreover, ncRNA expression may confer clinical information about disease outcomes and have utility as diagnostic tests[9,14]. The characterization of RNA species, their functions and their clinical applicability is therefore a major area of biological and clinical importance.

Here, we describe a comprehensive analysis of lincRNAs in 102 prostate cancer tissue samples and cell lines by RNA-Seq. We apply *ab initio* computational approaches to delineate the annotated and unannotated transcripts in this disease, and we find 121 ncRNAs, termed PCATs, whose expression patterns distinguish benign, localized cancer and metastatic cancer samples. Notably, we discover *PCAT-1*, a previously undescribed prostate cancer ncRNA that demonstrates either repression by PRC2 or an active role in promoting cell proliferation through transcriptional regulation of target genes. To our knowledge, our findings describe the first comprehensive study of lincRNAs in prostate cancer, provide a computational framework for large-scale RNA-Seq analyses and describe *PCAT-1* as a prostate cancer ncRNA functionally implicated in disease progression.

RESULTS

RNA-Seq analysis of the prostate cancer transcriptome

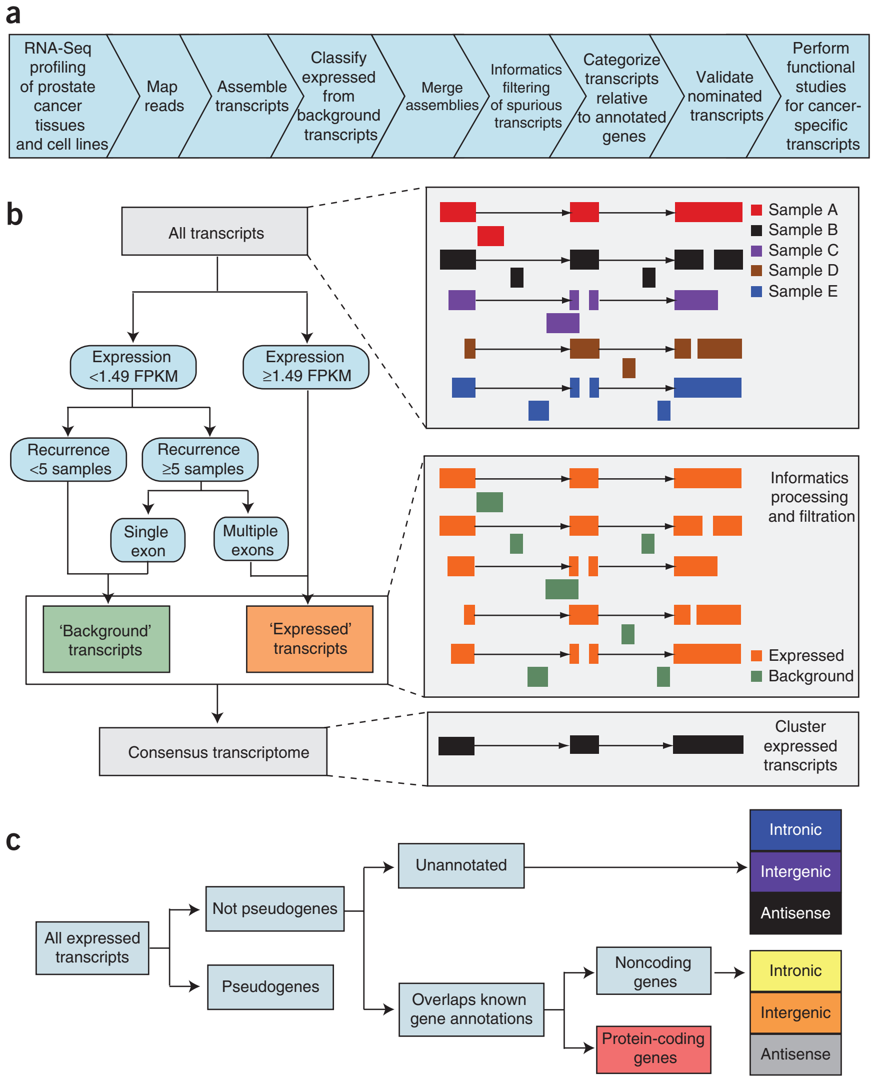
Over two decades of research have generated a genetic model of prostate cancer based on numerous neoplastic events, such as loss of the *PTEN*[15] tumor suppressor gene and gain of oncogenic ETS family transcription factor gene fusions[16–18] in large subsets of prostate cancer patients. As some patients lack these genetic aberrations, we hypothesized that prostate cancer similarly harbored disease-associated ncRNAs that characterized specific molecular subtypes.

To pursue this hypothesis, we applied transcriptome sequencing on a cohort of 102 prostate tissues and cell lines—20 benign adjacent prostates (benign), 47 localized prostate cancers (PCA), 14 metastatic

[1]Michigan Center for Translational Pathology, University of Michigan Medical School, Ann Arbor, Michigan, USA. [2]Department of Pathology, University of Michigan Medical School, Ann Arbor, Michigan, USA. [3]Department of Medicine, University of Chicago, Chicago, Illinois, USA. [4]Department of Urology, University of Michigan Medical School, Ann Arbor, Michigan, USA. [5]Department of Statistics, Colorado State University, Fort Collins, Colorado, USA. [6]Comprehensive Cancer Center, University of Michigan Medical School, Ann Arbor, Michigan, USA. [7]Howard Hughes Medical Institute, University of Michigan Medical School, Ann Arbor, Michigan, USA. [8]These authors contributed equally to this work. Correspondence should be addressed to A.M.C. (arul@umich.edu).

Received 11 January; accepted 10 June; published online 31 July 2011; doi:10.1038/nbt.1914

Figure 1 Analysis of transcriptome data for the detection of unannotated transcripts. (**a**) Schematic overview of the methodology employed in this study. (**b**) Graphical representation of the bioinformatics filters used to merge individual transcriptome libraries into a single consensus transcriptome. The merged consensus transcriptome was generated by compiling all individual transcriptome libraries and using individual decision tree classifiers for each chromosome to define high-confidence 'expressed' transcripts and low-confidence 'background' transcripts, which were discarded. The example decision tree on the left was trained on transcripts on chromosome 1. The graphics on the right illustrate the application of the informatics filtration pipeline to sample assembled transcripts. (**c**) After informatic processing and filtration of the sequencing data, transcripts were categorized to identify unannotated ncRNAs. Transcribed pseudogenes were isolated, and the remaining transcripts were categorized based on overlap with an aggregated set of known gene annotations into annotated protein coding, noncoding and unannotated. Both annotated and unannotated ncRNA transcripts were then separated into intronic, intergenic and antisense categories based on their relationship to protein-coding genes.



tumors and 21 prostate cell lines. From a total of 1.723 billion sequence fragments from 201 lanes of sequencing (108 paired-end and 93 single reads on the Illumina Genome Analyzer and Genome Analyzer II), we performed short-read gapped alignment[19] and recovered 1.41 billion mapped reads, with a median of 14.7 million mapped reads per sample (**Supplementary Table 1**). We used the Cufflinks *ab initio* assembly approach[3] to produce, for each sample, the most probable set of putative transcripts that served as the RNA templates for the sequence fragments in that sample (**Fig. 1a** and **Supplementary Figs. 1** and **2**).

As expected from a large tumor tissue cohort, individual transcript assemblies may have sources of noise, such as artifacts of the sequence alignment process, unspliced intronic pre-mRNA and genomic DNA contamination. To exclude these from our analyses, we trained a decision tree to classify transcripts as expressed versus background on the basis of transcript length, number of exons, recurrence in multiple samples and other structural characteristics (**Fig. 1b**, left, and **Supplementary Methods**). The classifier demonstrated a sensitivity of 70.8% and specificity of 88.3% when trained using transcripts that overlapped genes in the AceView database[20], including 11.7% of unannotated transcripts that were classified as expressed (**Fig. 1b** right). We then clustered the expressed transcripts into a consensus transcriptome and applied additional heuristic filters to further refine the assembly (**Supplementary Methods**). The final *ab initio* transcriptome assembly yielded 35,415 distinct transcriptional loci (**Supplementary Table 2** and **Supplementary Methods**).

Discovery of prostate cancer noncoding RNAs

We compared the assembled prostate cancer transcriptome to the UCSC, Ensembl, RefSeq, Vega and ENCODE gene databases to identify and categorize transcripts (**Fig. 1c**). The majority of the transcripts (77.3%) corresponded to annotated protein coding genes (72.1%) and noncoding RNAs (5.2%), but a substantial percentage (19.8%) lacked any overlap and were designated unannotated (**Fig. 2a**). These included partially intronic antisense (2.44%), totally intronic (12.1%) and intergenic transcripts (5.25%), consistent with previous reports of unannotated transcription[21–23]. Because of the added complexity of characterizing antisense or partially intronic transcripts without strand-specific RNA-Seq libraries, we focused on totally intronic and intergenic transcripts.

Global characterization of unannotated intronic and intergenic transcripts demonstrated that they were more highly expressed (**Fig. 2b**), had greater overlap with expressed sequence tags (ESTs) (**Supplementary Fig. 3**) and displayed a clear but subtle increase in conservation over randomly permuted controls (intergenic transcripts $P = 2.7 \times 10^{-4} \pm 0.0002$ for $0.4 < \omega < 0.8$; intronic transcripts $P = 2.6 \times 10^{-5} \pm 0.0017$ for $0 < \omega < 0.4$, Fisher's exact test, **Fig. 2c**). By contrast, unannotated transcripts scored lower than protein-coding genes for these metrics, which corroborates data in previous reports[2,24]. Notably, a small subset of unannotated intronic transcripts showed a profound degree of conservation (**Fig. 2c**, inset). Finally, analysis of coding potential revealed that only 5 of 6,144 transcripts harbored a high-quality open reading frame (ORF), indicating that the vast majority of these transcripts represent ncRNAs (**Supplementary Fig. 4**).

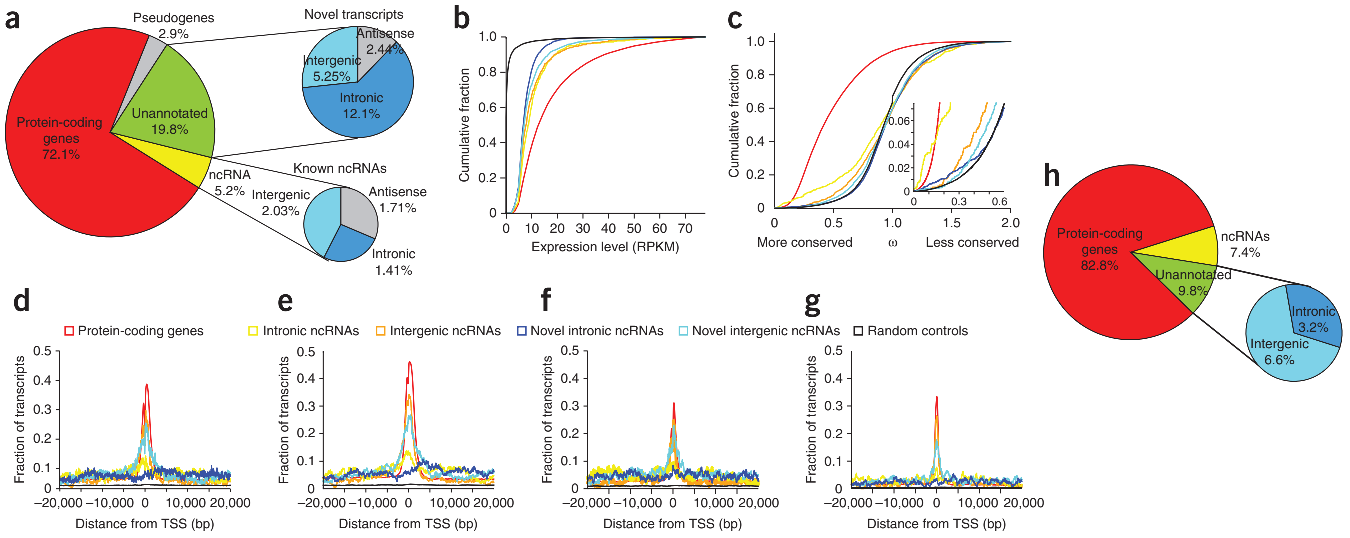


Figure 2 Prostate cancer transcriptome sequencing reveals dysregulation of unannotated transcripts. (**a**) Global overview of transcription in prostate cancer. The pie chart on the left displays transcript distribution in prostate cancer. The pie charts on the right display unannotated (upper) or annotated (lower) ncRNAs categorized as sense transcripts (intergenic and intronic) and antisense transcripts, respectively. (**b**) Line graph showing that unannotated transcripts are more highly expressed (reads per kilobase of transcript per million mapped reads; RPKM) than control regions. Negative control intervals were generated by randomly permuting the genomic positions of the transcripts. (**c**) Conservation analysis comparing unannotated transcripts to known genes and intronic controls shows a subtle degree of purifying selection among unannotated transcripts. The inset on the right shows an enlarged view. (**d**–**g**) Intersection plots displaying the fraction of unannotated transcripts enriched for H3K4me2 (**d**), H3K4me3 (**e**), acetyl-H3 (**f**) or RNA polymerase II (**g**) at their transcriptional start site (TSS) using ChIP-Seq and RNA-Seq data for the VCaP prostate cancer cell line. The legend applies to plots in **b**–**g**. (**h**) A pie chart displaying the distribution of differentially expressed transcripts in prostate cancer (FDR < 0.01).

To determine whether our unannotated transcripts were supported by histone modifications defining active transcriptional units, we used published prostate cancer chromatin immunoprecipitation (ChIP)-Seq data for two prostate cell lines[25], VCaP and LNCaP (**Supplementary Table 3**). After filtering our data set for transcribed repetitive elements known to display alternative patterns of histone modifications[26], we observed a strong enrichment for histone modifications characterizing transcriptional start sites (TSSs) and active transcription, including H3K4me2, H3K4me3, acetyl-H3 and RNA polymerase II (**Fig. 2d**–**g**), but not H3K4me1, which characterizes enhancer regions[27] (**Supplementary Figs. 5** and **6**). Notably, intergenic ncRNAs showed greater enrichment compared to intronic ncRNAs in these analyses (**Fig. 2d**–**g**).

To elucidate global changes in transcript abundance in prostate cancer, we analyzed differential expression for all transcripts. We found 836 genes differentially expressed between benign samples and localized tumors (false-discovery rate (FDR) < 0.01), with annotated protein-coding and ncRNA genes constituting 82.8% and 7.4% of differentially expressed genes, respectively, including known prostate cancer biomarkers such *AMACR*[28], *HPN*[29] and *PCA3* (ref. 14) (**Fig. 2h**, **Supplementary Fig. 2** and **Supplementary Table 4**). Finally, 9.8% of differentially expressed genes corresponded to unannotated ncRNAs, including 3.2% within gene introns and 6.6% in intergenic regions.

Characterization of PCATs

As ncRNAs may contribute to human disease[6–9], we identified aberrantly expressed uncharacterized ncRNAs in prostate cancer. We found a total of 1,859 unannotated lincRNAs throughout the human genome. Overall, these intergenic RNAs resided approximately halfway between two protein coding genes (**Supplementary Fig. 7**), and over one-third (34.1%) were ≥10 kb from the nearest protein-coding gene, which is consistent with previous reports[30] and supports the independence of intergenic ncRNAs genes. For example, visualizing the Chr15q arm using the Circos program (http://circos.ca/) illustrated genomic positions of 89 unannotated intergenic transcripts, including one differentially expressed gene centromeric to *TLE3* (**Supplementary Fig. 8**).

A focused analysis of the 1,859 unannotated intergenic RNAs yielded 106 that were differentially expressed in localized tumors (FDR < 0.05, **Fig. 3a**). A cancer outlier expression analysis (**Supplementary Methods**) similarly nominated numerous unannotated ncRNA outliers (**Fig. 3b**) as well as known prostate cancer outliers, such as *ERG*[18], *ETV1* (refs. 17,18), *SPINK1* (ref. 31) and *CRISP3* (ref. 32). Merging these results produced a set of 121 unannotated transcripts that accurately discriminated benign, localized tumor and metastatic prostate samples by unsupervised clustering (**Fig. 3a**). Indeed, clustering analyses using unannotated ncRNA outliers also suggested disease subtypes (**Supplementary Fig. 9**). These 121 unannotated transcripts were ranked and named as PCATs according to their fold-change in localized tumor versus benign tissue (**Supplementary Tables 5**–**7**).

Validation of novel ncRNAs

To gain confidence in our transcript nominations, we validated multiple unannotated transcripts *in vitro* by reverse transcription PCR (RT-PCR) and quantitative real-time PCR (qPCR) (**Supplementary Fig. 10**). qPCR for four transcripts (*PCAT-114*, *PCAT-14*, *PCAT-43* and *PCAT-1*) on two independent cohorts of prostate tissues confirmed predicted cancer-specific expression patterns (**Fig. 3c**–**f** and **Supplementary Fig. 11**). Notably, all four are prostate-specific, with minimal expression seen by qPCR in breast (n = 14) or lung cancer (n = 16) cell lines or in 19 normal tissue types (**Supplementary Table 8**). This is further supported by expression analysis of these transcripts in

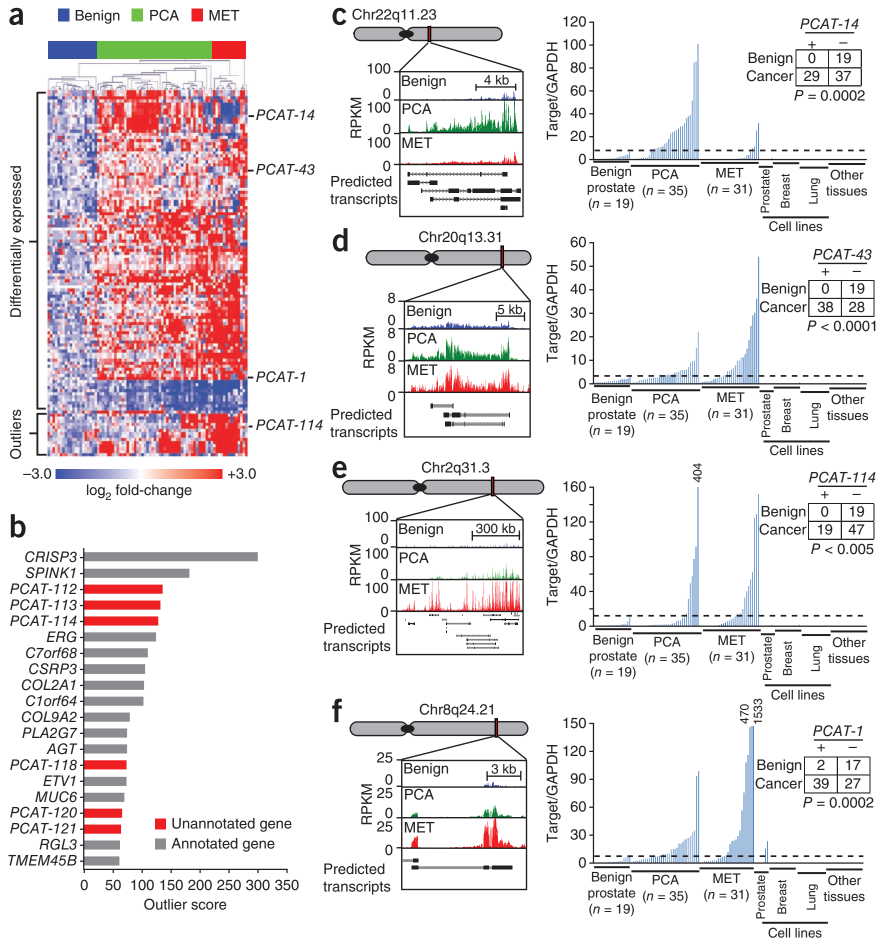


Figure 3 Unannotated intergenic transcripts differentiate prostate cancer and benign prostate samples. (**a**) Unsupervised clustering analyses of differentially expressed or outlier unannotated intergenic transcripts clusters benign samples, localized tumors and metastatic cancers. Expression is plotted as $\log_2$ fold-change relative to the median of the benign samples. The four transcripts detailed in this study are indicated on the side. (**b**) Cancer outlier expression analysis for the prostate cancer transcriptome ranks unannotated transcripts prominently. (**c**–**f**) qPCR on an independent cohort of prostate and nonprostate samples (benign ($n = 19$), PCA ($n = 35$), metastatic (MET) ($n = 31$), prostate cell lines ($n = 7$), breast cell lines ($n = 14$), lung cell lines ($n = 16$), other normal samples ($n = 19$); **Supplementary Table 8**)) measures expression levels of four nominated ncRNAs—*PCAT-14* (**c**), *PCAT-43* (**d**), *PCAT-114* (**e**), *PCAT-1* (**f**)—and upregulated in prostate cancer. Inset tables on the right quantify 'positive' and 'negative' expressing samples using the cut-off value (shown as a black dashed lines). Statistical significance was determined using a Fisher's exact test. qPCR analysis was performed by normalizing to *GAPDH* and the median expression of the benign samples.

our RNA-Seq compendium of 13 tumor types, representing 325 samples (**Supplementary Fig. 12**). This tissue specificity was not necessarily due to regulation by androgen receptor signaling, as only *PCAT-14* expression was induced when androgen responsive VCaP and LNCaP cells were treated with the synthetic androgen R1881, consistent with previous data from this locus[17] (**Supplementary Fig. 13**). *PCAT-1* and *PCAT-14* also showed cancer-specific upregulation when tested on a panel of matched tumor-normal pair samples (**Supplementary Fig. 14**).

Of note, *PCAT-114*, which ranks as the fifth best outlier, just ahead of *ERG* (**Fig. 3b** and **Supplementary Table 7**), appears as part of a large, >500 kb locus of expression in a gene desert in Chr2q31. We termed this region 'second chromosome locus associated with prostate-1' (SChLAP1) (**Supplementary Fig. 15**). Careful analysis of the SChLAP1 locus revealed both discrete transcripts and intronic transcription, highlighting this region as an intriguing aspect of the prostate cancer transcriptome.

PCAT-1, an unannotated prostate cancer lincRNA

To explore several transcripts more closely, we carried out 5′ and 3′ rapid amplification of cDNA ends (RACE) for *PCAT-1* and *PCAT-14*. Interestingly, the *PCAT-14* locus contained components of viral ORFs from the HERV-K endogenous retrovirus family (**Supplementary Fig. 16**), whereas *PCAT-1* incorporates portions of a mariner family transposase[33,34], an *Alu* and a viral long terminal repeat promoter region (**Fig. 4a** and **Supplementary Fig. 17**). Whereas *PCAT-14* was upregulated in localized prostate cancer but largely absent in metastases (**Fig. 3c**), *PCAT-1* was strikingly upregulated in a subset of metastatic and high-grade localized (Gleason score ≥7) cancers (**Fig. 3f** and **Supplementary Fig. 11**). Because of this notable profile, we hypothesized that *PCAT-1* may have coordinated expression with the oncoprotein *EZH2*, a core PRC2 protein that is upregulated in solid tumors and contributes to a metastatic phenotype[35,36]. Surprisingly, we found that *PCAT-1* and *EZH2* expression were nearly mutually exclusive (**Fig. 4b**), with only one patient showing outlier expression of both. This suggests that outlier *PCAT-1* and *EZH2* expression may define two subsets of high-grade disease.

PCAT-1 is located in the chromosome 8q24 gene desert ~725 kb upstream of the *c-MYC* oncogene. To confirm that *PCAT-1* is a noncoding gene, we cloned the full-length *PCAT-1* transcript and performed *in vitro* translational assays, which were negative as expected (**Supplementary Fig. 18**). Next, because Chr8q24 is known to harbor prostate cancer–associated single nucleotide polymorphisms (SNPs) and to exhibit frequent chromosomal amplification[37–42], we evaluated whether the relationship between *EZH2* and *PCAT-1* was specific or generalized. To address this, we measured expression levels of *c-MYC* and *NCOA2*, two proposed targets of Chr8q amplification[39,42], by qPCR. Neither *c-MYC* nor *NCOA2* levels showed striking expression relationships to *PCAT-1*, *EZH2* or each other (**Supplementary Fig. 19**). Likewise, *PCAT-1* outlier expression was not dependent on Chr8q24 amplification, as highly expressing localized tumors often did not have 8q24 amplification and high copy number gain of 8q24 was not sufficient to upregulate *PCAT-1* (**Supplementary Figs. 20** and **21**).

PCAT-1 function and regulation

Despite reports showing that upregulation of the ncRNA *HOTAIR* participates in PRC2 function in breast cancer[9], we do not observe strong expression of this ncRNA in prostate (**Supplementary Fig. 22**), suggesting that other ncRNAs may be important in this cancer. To determine the mechanism for the expression profiles of *PCAT-1* and

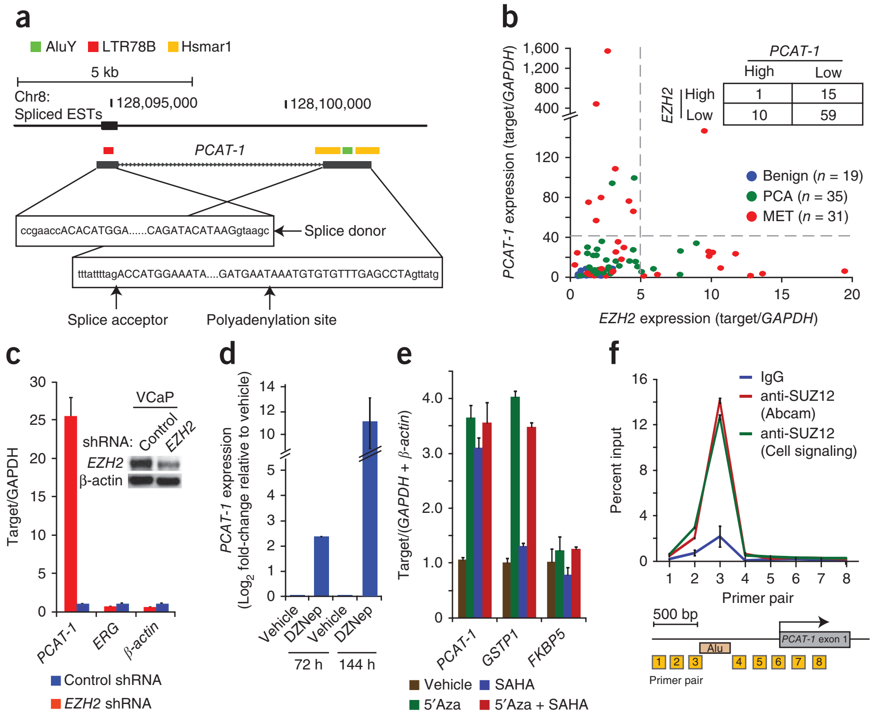


Figure 4 *PCAT-1* is a marker of aggressive cancer and a PRC2-repressed ncRNA. (**a**) The genomic location of *PCAT-1* determined by 5′ and 3′ RACE, with DNA sequence features indicated by the colored boxes. (**b**) qPCR for *PCAT-1* (*y* axis) and *EZH2* (*x* axis) on a cohort of benign (*n* = 19), localized tumor (*n* = 35) and metastatic cancer (*n* = 31) samples. The inset table quantifies patient subsets demarcated by the gray dashed lines. (**c**) Knockdown of *EZH2* in VCaP resulted in upregulation of *PCAT-1*. Data were normalized to *GAPDH* and represented as fold-change. *ERG* and *B-actin* serve as negative controls. The inset western blot indicates *EZH2* knockdown. (**d**) Treatment of VCaP cells with 0.1 μM of the *EZH2* inhibitor DZNep or vehicle control (DMSO) shows increased expression of *PCAT-1* transcript after *EZH2* inhibition. (**e**) *PCAT-1* expression is increased upon treatment of VCaP cells with the demethylating agent 5′azacytidine (5′Aza), the histone deacetylase inhibitor SAHA or a combination of both. qPCR data were normalized to the average of (*GAPDH* + *β-actin*) and represented as fold-change. *GSTP1* and *FKBP5* are positive and negative controls, respectively. (**f**) ChIP assays for SUZ12 demonstrated direct binding of SUZ12 to the *PCAT-1* promoter. Primer locations are indicated (boxed numbers) in the *PCAT-1* schematic.

EZH2, we inhibited *EZH2* activity in VCaP cells, which express low-to-moderate levels of *PCAT-1*. Knockdown of *EZH2* by short hairpin (sh)RNA or pharmacologic inhibition of *EZH2* with the inhibitor 3-deazaneplanocin A (DZNep) caused a dramatic upregulation in *PCAT-1* expression levels (**Fig. 4c**,**d**), as did treatment of VCaP cells with the demethylating agent 5′deoxyazacytidine, the histone deacetylase inhibitor SAHA or both (**Fig. 4e**). ChIP assays also demonstrated that SUZ12, a core PRC2 protein, directly binds the *PCAT-1* promoter ~1 kb upstream of the TSS (**Fig. 4f**). Notably, RNA immunoprecipitation similarly showed binding of *PCAT-1* to SUZ12 protein in VCaP cells (**Supplementary Fig. 23a**). RNA immunoprecipitation assays followed by RNase A, RNase H or DNase I treatment either abolished, partially preserved or totally preserved this interaction, respectively (**Supplementary Fig. 23b**). This suggests that *PCAT-1* exists primarily as a single-stranded RNA and secondarily as a RNA/DNA hybrid.

To explore the functional role of *PCAT-1* in prostate cancer, we stably overexpressed full-length *PCAT-1* or controls in RWPE benign immortalized prostate cells. We observed a modest but consistent increase in cell proliferation when *PCAT-1* was overexpressed at physiological levels (**Fig. 5a** and **Supplementary Fig. 24**). Next, we designed short interfering (si)RNA oligos to *PCAT-1* and performed knockdown experiments in LNCaP cells, which express higher levels of *PCAT-1* without PRC2-mediated repression (**Supplementary Fig. 25**). Supporting our overexpression data, knockdown of *PCAT-1* with three independent siRNA oligos resulted in a 25–50% decrease in cell proliferation in LNCaP cells (**Fig. 5b**), but not in control DU145 cells lacking *PCAT-1* expression (**Supplementary Fig. 26**) or VCaP cells, in which *PCAT-1* is expressed but repressed by PRC2 (**Supplementary Fig. 27**).

Gene expression profiling of LNCaP knockdown samples on cDNA microarrays indicated that *PCAT-1* modulates the transcriptional regulation of 370 genes (255 upregulated, 115 downregulated; FDR ≤ 0.01) (**Supplementary Fig. 28** and **Supplementary Table 9**). Gene ontology analysis of the upregulated genes showed preferential enrichment for gene set concepts such as mitosis and cell cycle, whereas the downregulated genes had no concepts showing statistical significance (**Fig. 5c** and **Supplementary Table 10**). These results suggest that the function of *PCAT-1* is predominantly repressive in nature, similar to other lincRNAs. We next validated expression

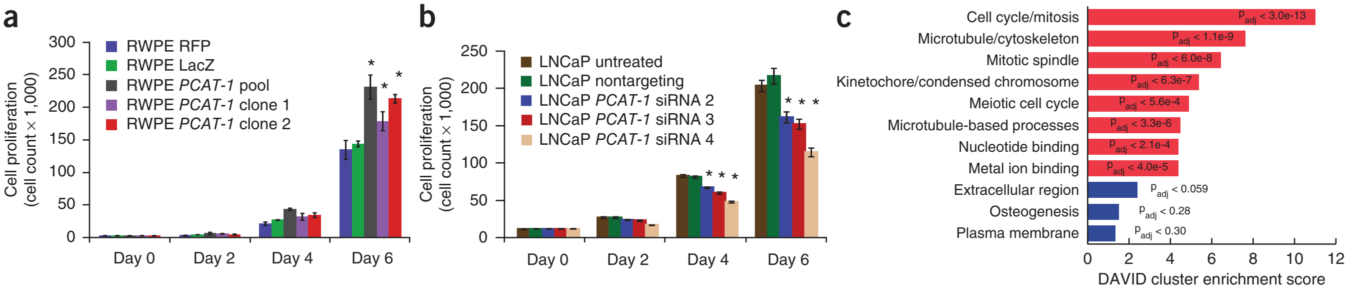


Figure 5 *PCAT-1* promotes cell proliferation. (**a**) Cell proliferation assays for RWPE benign immortalized prostate cells stably infected with *PCAT-1* lentivirus or RFP and LacZ control lentiviruses. An asterisk (*) indicates $P \leq 0.02$ by a two-tailed Student's *t*-test. (**b**) Cell proliferation assays in LNCaP using *PCAT-1* siRNAs. An asterisk (*) indicates $P \leq 0.005$ by a two-tailed Student's *t*-test. (**c**) Gene ontology analysis of *PCAT-1* knockdown microarray data using the DAVID program. Blue bars represent the top hits for upregulated genes. Red bars represent the top hits for downregulated genes. DAVID enrichment scores are represented with Benjamini-Hochberg-adjusted *P* values. All error bars in this figure are mean ± s.e.m.

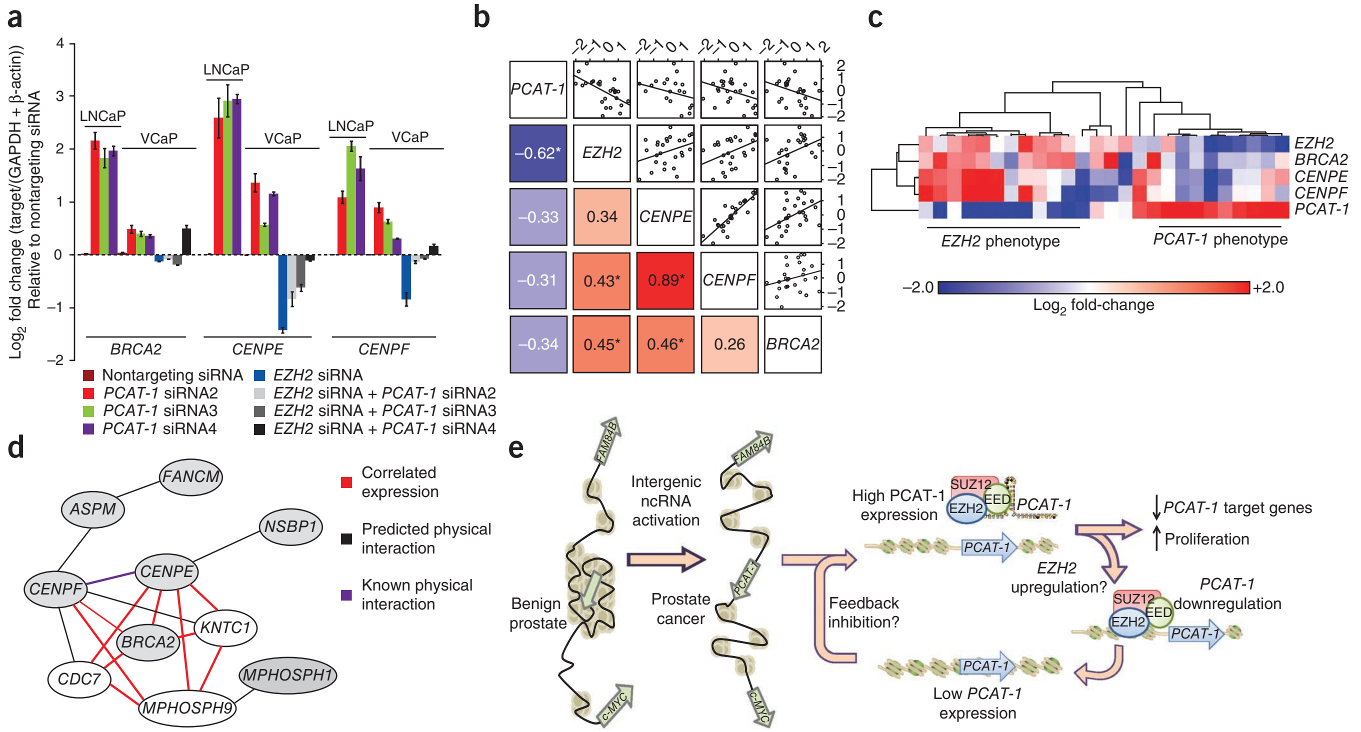


Figure 6 Prostate cancer tissues recapitulate *PCAT-1* signaling. (**a**) qPCR expression of three *PCAT-1* target genes after *PCAT-1* knockdown in VCaP and LNCaP cells, as well as following *EZH2* knockdown or dual *EZH2* and *PCAT-1* knockdown in VCaP cells. qPCR data were normalized to the average of (*GAPDH* + *β-actin*) and represented as fold change. Error bars represent mean ± s.e.m. (**b**) Standardized log2-transformed qPCR expression of a set of tumors and metastases with outlier expression of either *PCAT-1* or *EZH2*. The shaded squares in the lower left show Spearman correlation values between the indicated genes (* indicates $P < 0.05$). Blue and red indicate negative or positive correlation, respectively. The upper squares show the scatter plot matrix and fitted trend lines for the same comparisons. (**c**) A heatmap of *PCAT-1* target genes (*BRCA2*, *CENPF*, *CENPE*) in *EZH2*-outlier and *PCAT-1*-outlier patient samples (see **Fig. 4b**). Expression was determined by qPCR and normalized as in **b**. (**d**) A predicted network generated by the HefaLMP program for 7 of 20 top upregulated genes following *PCAT-1* knockdown in LNCaP cells. Gray nodes are genes found following *PCAT-1* knockdown. Red edges indicate co-expressed genes; black edges indicate predicted protein-protein interactions; and purple edges indicate verified protein-protein interactions. (**e**) A proposed schematic representing *PCAT-1* upregulation, function and relationship to PRC2.

changes in three key *PCAT-1* target genes (*BRCA2*, *CENPE* and *CENPF*) whose expression is upregulated upon *PCAT-1* knockdown (**Fig. 6a**) in LNCaP and VCaP cells, the latter of which appear less sensitive to *PCAT-1* knockdown likely due to lower overall expression levels of this transcript.

PCAT-1 signatures in prostate cancer

Because of the regulation of *PCAT-1* by PRC2 in VCaP cells, we hypothesized that knockdown of *EZH2* would also downregulate *PCAT-1* targets as a secondary phenomenon owing to the subsequent upregulation of *PCAT-1*. Simultaneous knockdown of *PCAT-1* and *EZH2* would thus abrogate expression changes in *PCAT-1* target genes. Carrying out this experiment in VCaP cells demonstrated that *PCAT-1* target genes were indeed downregulated by *EZH2* knockdown, and that this change was either partially or completely reversed using siRNA oligos to *PCAT-1* (**Fig. 6a**), lending support to the role of *PCAT-1* as a transcriptional repressor. Taken together, these results suggest that *PCAT-1* biology may exhibit two distinct modalities: one in which PRC2 represses *PCAT-1* and a second in which active *PCAT-1* promotes cell proliferation. *PCAT-1* and PRC2 may therefore characterize distinct subsets of prostate cancer.

To examine these findings, we used qPCR to measure expression of *BRCA2*, *CENPE* and *CENPF* in our cohort of tissue samples. Consistent with our model, we found that samples expressing *PCAT-1* tended to have low expression of *PCAT-1* target genes (**Fig. 6b**). Moreover, comparing *EZH2*-outlier and *PCAT-1*-outlier patients (**Fig. 4b**), we found that two distinct phenotypes emerged. Individuals with high *EZH2* tended to have high levels of *PCAT-1* target genes, and those with high expression of *PCAT-1* itself displayed the opposite expression pattern of target genes (**Fig. 6c**). Network analysis of the top 20 upregulated genes after *PCAT-1* knockdown with the HefaLMP tool[43] further suggested that these genes form a coordinated network (**Fig. 6d**), corroborating our previous observations. Taken together, these results provide initial data into the composition and function of the prostate cancer ncRNA transcriptome.

DISCUSSION

To our knowledge, this study represents the largest RNA-Seq analysis to date and the first to comprehensively analyze a common epithelial cancer from a large cohort of human tissue samples. As such, our study has adapted existing computational tools intended for small-scale use[3] and developed new methods to distill large numbers of transcriptome data sets into a single consensus transcriptome assembly that accurately represents disease biology (**Supplementary Discussion**).

Among the numerous uncharacterized ncRNA species detected by our study, we have focused on 121 PCATs, which we believe represent a set of uncharacterized ncRNAs that may have important biological functions in this disease. In this regard, these data contribute to a growing body of literature supporting the importance of unannotated ncRNA species in cellular biology and oncogenesis[6–12], and broadly our study confirms the utility of RNA-Seq in defining functionally important elements of the genome[2–4].

Of particular interest is our discovery of the prostate-specific ncRNA gene *PCAT-1*, which is markedly overexpressed in a subset of prostate cancers, particularly metastases, and may contribute to cell proliferation in these tumors. It is also notable that *PCAT-1* resides in the 8q24 'gene desert' locus, in the vicinity of well-studied prostate cancer risk SNPs and the *c-MYC* oncogene, suggesting that this locus—and its frequent amplification in cancer—may be linked to additional aspects of cancer biology (**Supplementary Discussion**). In addition, the interplay between PRC2 and *PCAT-1* further suggests that this ncRNA may have an important role in prostate cancer progression (**Fig. 6e**). Other ncRNAs identified by this analysis may similarly contribute to prostate cancer as well. Furthermore, recent preclinical efforts to detect prostate cancer noninvasively through the collection of patient urine samples have shown promise for several urine-based prostate cancer biomarkers, including the ncRNA *PCA3* (refs. 44,45). Although additional studies are needed, our identification of ncRNA biomarkers for prostate cancer suggests that urine-based assays for these ncRNAs may also warrant investigation, particularly for those that may stratify patient molecular subtypes.

Our findings support an important role for tissue-specific ncRNAs in prostate cancer and suggest that cancer-specific functions of these ncRNAs may help to drive tumorigenesis. We further speculate that specific ncRNA signatures may occur universally in all disease states and that applying these methodologies to other diseases may reveal key aspects of disease biology and clinically important biomarkers.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturebiotechnology/.

Accession codes. Data from RNA-Seq experiments are deposited at the NCBI Gene Expression Omnibus as GSE25183. *PCAT-1* and *PCAT-14* nucleotide sequences are deposited at GenBank nucleotide database (nuccore) as HQ605084 and HQ605085, respectively.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

We thank K. Ramnarayanan and R. Morey for technical assistance with next generation sequencing. We thank R.J. Lonigro, S. Kaylana-Sundaram, T. Barrette, and M. Quist for help with sequencing data analysis, and R. Mehra, B. Han and K. Suleman for prostate tissue specimens. We thank C. Trapnell and G. Pertea for assistance with computational analyses. We thank S. Tomlins, Y.-M. Wu, S. Roychowdhury and members of the Chinnaiyan laboratory for advice and discussions. We thank R. Beroukhim for guidance. This work was supported in part by the US National Institutes of Health (NIH) Prostate Specialized Program of Research Excellence grant P50CA69568, the Early Detection Research Network grant UO1 CA111275 (to A.M.C.), the NIH R01CA132874-01A1 (to A.M.C.), the Department of Defense grant W81XWH-10-0652 and W81XWH-11-1-0337 (to A.M.C.) and the National Center for Functional Genomics supported by the Department of Defense (to A.M.C.). A.M.C. is supported by the Doris Duke Charitable Foundation Clinical Scientist Award, a Burroughs Wellcome Foundation Award in Clinical Translational Research and the Prostate Cancer Foundation. A.M.C. is an American Cancer Society Research Professor. N.P. was supported by a University of Michigan Prostate SPORE Career Development Award. C.A.M. was supported by the American Association of Cancer Research Amgen Fellowship in Clinical/Translational Research, the Canary Foundation and American Cancer Society Early Detection Postdoctoral Fellowship, and a Prostate Cancer Foundation Young Investigator Award. Q.C. was supported by a Department of Defense Postdoctoral Fellowship grant PC094725. J.R.P. was supported by the NIH Cancer Biology Training grant CA009676-18 and the Department of Defense Predoctoral Fellowship W81XWH-10-1-0551. M.K.I. was supported by the Department of Defense Predoctoral Fellowship W81XWH-11-1-0136. J.R.P. and M.K.I. are Fellows of the University of Michigan Medical Scientist Training Program.

AUTHOR CONTRIBUTIONS

M.K.I., J.R.P. and A.M.C. designed the project and directed experimental studies. M.K.I., O.A.B., C.S.G. and C.A.M. developed computational platforms and performed sequencing data analysis. M.K.I., O.A.B. and H.K.I. performed statistical analyses. J.R.P., S.M.D., J.C.B., Q.C., N.P., H.D.K., B.L., X.W., I.A.A., X.C., X.J. and D.R. performed experimental studies. J.S. and J.T.W. coordinated biospecimens. M.K.I., J.R.P. and A.M.C. interpreted data and wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturebiotechnology/.

Published online at http://www.nature.com/naturebiotechnology/.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Metzker, M.L. Sequencing technologies—the next generation. *Nat. Rev. Genet.* **11**, 31–46 (2010).
2. Guttman, M. *et al. Ab initio* reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat. Biotechnol.* **28**, 503–510 (2010).
3. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
4. Robertson, G. *et al. De novo* assembly and analysis of RNA-seq data. *Nat. Methods* **7**, 909–912 (2010).
5. Zerbino, D.R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18**, 821–829 (2008).
6. Huarte, M. *et al.* A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. *Cell* **142**, 409–419 (2010).
7. Orom, U.A. *et al.* Long noncoding RNAs with enhancer-like function in human cells. *Cell* **143**, 46–58 (2010).
8. Rinn, J.L. *et al.* Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* **129**, 1311–1323 (2007).
9. Gupta, R.A. *et al.* Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* **464**, 1071–1076 (2010).
10. Pasmant, E. *et al.* Characterization of a germ-line deletion, including the entire INK4/ARF locus, in a melanoma-neural system tumor family: identification of ANRIL, an antisense noncoding RNA whose expression coclusters with ARF. *Cancer Res.* **67**, 3963–3969 (2007).
11. Yap, K.L. *et al.* Molecular interplay of the noncoding RNA ANRIL and methylated histone H3 lysine 27 by polycomb CBX7 in transcriptional silencing of INK4a. *Mol. Cell* **38**, 662–674 (2010).
12. Tsai, M.C. *et al.* Long noncoding RNA as modular scaffold of histone modification complexes. *Science* **329**, 689–693 (2010).
13. Kotake, Y. *et al.* Long non-coding RNA ANRIL is required for the PRC2 recruitment to and silencing of p15(INK4B) tumor suppressor gene. *Oncogene* **30**, 1956–1962 (2011).
14. de Kok, J.B. *et al.* DD3(PCA3), a very sensitive and specific marker to detect prostate tumors. *Cancer Res.* **62**, 2695–2698 (2002).
15. Li, J. *et al.* PTEN, a putative protein tyrosine phosphatase gene mutated in human brain, breast, and prostate cancer. *Science* **275**, 1943–1947 (1997).
16. Prensner, J.R. & Chinnaiyan, A.M. Oncogenic gene fusions in epithelial carcinomas. *Curr. Opin. Genet. Dev.* **19**, 82–91 (2009).
17. Tomlins, S.A. *et al.* Distinct classes of chromosomal rearrangements create oncogenic ETS gene fusions in prostate cancer. *Nature* **448**, 595–599 (2007).
18. Tomlins, S.A. *et al.* Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* **310**, 644–648 (2005).
19. Trapnell, C., Pachter, L. & Salzberg, S.L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).
20. Thierry-Mieg, D. & Thierry-Mieg, J. AceView: a comprehensive cDNA-supported gene and transcripts annotation. *Genome Biol.* **7** (suppl. 1), S11–S14 (2006).
21. Birney, E. *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816 (2007).
22. The FANTOM Consortium. The transcriptional landscape of the mammalian genome. *Science* **309**, 1559–1563 (2005).
23. He, Y., Vogelstein, B., Velculescu, V.E., Papadopoulos, N. & Kinzler, K.W. The antisense transcriptomes of human cells. *Science* **322**, 1855–1857 (2008).
24. Guttman, M. *et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**, 223–227 (2009).
25. Yu, J. *et al.* An integrated network of androgen receptor, polycomb, and TMPRSS2-ERG gene fusions in prostate cancer progression. *Cancer Cell* **17**, 443–454 (2010).
26. Day, D.S., Luquette, L.J., Park, P.J. & Kharchenko, P.V. Estimating enrichment of repetitive elements from high-throughput sequence data. *Genome Biol.* **11**, R69 (2010).
27. Kim, T.K. *et al.* Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**, 182–187 (2010).
28. Rubin, M.A. *et al.* Alpha-methylacyl coenzyme A racemase as a tissue biomarker for prostate cancer. *J. Am. Med. Assoc.* **287**, 1662–1670 (2002).

29. Dhanasekaran, S.M. *et al.* Delineation of prognostic biomarkers in prostate cancer. *Nature* **412**, 822–826 (2001).
30. van Bakel, H., Nislow, C., Blencowe, B.J. & Hughes, T.R. Most "dark matter" transcripts are associated with known genes. *PLoS Biol.* **8**, e1000371 (2010).
31. Tomlins, S.A. *et al.* The role of SPINK1 in ETS rearrangement-negative prostate cancers. *Cancer Cell* **13**, 519–528 (2008).
32. Bjartell, A.S. *et al.* Association of cysteine-rich secretory protein 3 and beta-microseminoprotein with outcome after radical prostatectomy. *Clin. Cancer Res.* **13**, 4130–4138 (2007).
33. Oosumi, T., Belknap, W.R. & Garlick, B. Mariner transposons in humans. *Nature* **378**, 672 (1995).
34. Robertson, H.M., Zumpano, K.L., Lohe, A.R. & Hartl, D.L. Reconstructing the ancient mariners of humans. *Nat. Genet.* **12**, 360–361 (1996).
35. Kleer, C.G. *et al.* EZH2 is a marker of aggressive breast cancer and promotes neoplastic transformation of breast epithelial cells. *Proc. Natl. Acad. Sci. USA* **100**, 11606–11611 (2003).
36. Varambally, S. *et al.* The polycomb group protein EZH2 is involved in progression of prostate cancer. *Nature* **419**, 624–629 (2002).
37. Ahmadiyeh, N. *et al.* 8q24 prostate, breast, and colon cancer risk loci show tissue-specific long-range interaction with MYC. *Proc. Natl. Acad. Sci. USA* **107**, 9742–9746 (2010).
38. Al Olama, A.A. *et al.* Multiple loci on 8q24 associated with prostate cancer susceptibility. *Nat. Genet.* **41**, 1058–1060 (2009).
39. Beroukhim, R. *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature* **463**, 899–905 (2010).
40. Gudmundsson, J. *et al.* Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. *Nat. Genet.* **39**, 631–637 (2007).
41. Sotelo, J. *et al.* Long-range enhancers on 8q24 regulate c-Myc. *Proc. Natl. Acad. Sci. USA* **107**, 3001–3005 (2010).
42. Taylor, B.S. *et al.* Integrative genomic profiling of human prostate cancer. *Cancer Cell* **18**, 11–22 (2010).
43. Huttenhower, C. *et al.* Exploring the human genome with functional maps. *Genome Res.* **19**, 1093–1106 (2009).
44. Laxman, B. *et al.* A first-generation multiplex biomarker analysis of urine for the early detection of prostate cancer. *Cancer Res.* **68**, 645–649 (2008).
45. Hessels, D. *et al.* DD3(PCA3)-based molecular urine analysis for the diagnosis of prostate cancer. *Eur. Urol.* **44**, 8–16 (2003).

ONLINE METHODS

Cell lines, treatments and tissues. All prostate cell lines were obtained from the American Type Culture Collection, except for PrEC (benign nonimmortalized prostate epithelial cells) and PrSMC (prostate smooth muscle cells), which were obtained from Lonza. Cell lines were maintained using standard media and conditions.

For androgen treatment experiments, LNCaP and VCaP cells were grown in androgen-depleted media for 48 h and subsequently treated with 5nM methyltrienolone (R1881, NEN Life Science Products) or an equivalent volume of ethanol for 48 h before harvesting the cells. For drug treatments, VCaP cells were treated with 20 μM 5′deoxyazacytidine (Sigma), 500 nM HDAC inhibitor suberoylanilide hydroxamic acid (SAHA) (Biovision), or both 5′deoxyazacytidine and SAHA. 5′deoxyazacytidine treatments were performed for 6 d with media and drug reapplied every 48 h. SAHA treatments were done for 48 h. DMSO treatments were done for 6 d. For DZNep treatments, DZNep was dissolved in DMSO and VCaP cells were treated with either 0.1 μM of DZNep or vehicle control; RNA was harvested at 72 h and 144 h.

Prostate tissues were obtained from the radical prostatectomy series and Rapid Autopsy Program_ENREF_48 at the University of Michigan tissue core as part of the University of Michigan Prostate Cancer Specialized Program of Research Excellence (S.P.O.R.E.). All tissue samples were collected with informed consent under an Institutional Review Board (IRB) approved protocol at the University of Michigan.

RNA isolation, cDNA synthesis and PCR experiments. Total RNA was isolated using Trizol and an RNeasy Kit (Invitrogen) with DNase I digestion according to the manufacturer's instructions. RNA integrity was verified on an Agilent Bioanalyzer 2100 (Agilent Technologies). cDNA was synthesized from total RNA using Superscript III (Invitrogen) and random primers (Invitrogen). Quantitative Real-time PCR (qPCR) was done using Power SYBR Green Mastermix (Applied Biosystems) on an Applied Biosystems 7900HT Real-Time PCR System. (RT-PCR was done with Platinum Taq High Fidelity polymerase (Invitrogen). All oligonucleotide primers are listed in **Supplementary Table 11**. For PCR product sequencing, PCR products were resolved on a 1.5% agarose gel, and either sequenced directly or extracted using a Gel Extraction kit (Qiagen) and cloned into pcr4-TOPO vectors (Invitrogen). PCR products were bidirectionally sequenced at the University of Michigan Sequencing Core.

RNA-ligase–mediated rapid amplification of cDNA ends (RACE). 5′ and 3′ RACE was performed using the GeneRacer RLM-RACE kit (Invitrogen) according to the manufacturer's instructions. RACE PCR products were obtained using Platinum Taq high-fidelity polymerase (Invitrogen), the supplied GeneRacer primers, and appropriate gene-specific primers indicated in **Supplementary Table 11**.

RNA-Seq library preparation. 2 μg total RNA was selected for polyA$^+$ RNA using Sera-Mag oligo(dT) beads (Thermo Scientific), and paired-end next-generation sequencing libraries were prepared, as previously described[46], using Illumina-supplied universal adaptor oligos and PCR primers (Illumina). Samples were sequenced in a single lane on an Illumina Genome Analyzer I or Genome Analyzer II flow cell using previously described protocols[46]. 36–45 mer paired-end reads were done according to the protocol provided by Illumina.

Overexpression studies. *PCAT-1* full-length transcript was cloned into the pLenti6 vector (Invitrogen) along with RFP and LacZ controls. After confirmation of the insert sequence, lentiviruses were generated at the University of Michigan Vector Core and transfected into the benign immortalized prostate cell line RWPE. RWPE cells stably expressing *PCAT-1*, RFP or LacZ were generated by selection with blasticidin (Invitrogen), and 10,000 cells were plated into 12-well plates. Cells were harvested and counted at day 2, day 4 and day 6 post-plating with a Coulter counter.

siRNA knockdown studies. Cells were plated and transfected with 20 μM experimental siRNA oligos or nontargeting controls twice, at 12 h and 36 h post-plating. Knockdowns were performed with Oligofectamine in OptiMEM media. Knockdown efficiency was determined by qPCR. siRNA sequences (in sense format) for *PCAT-1* knockdown were as follows: siRNA 1 UUAAAGAGAUCCACAGUUAUU; siRNA 2 GCAGAAACACCAAUGGAUAUU; siRNA 3 AUACAUAAGACCAUGGAAAU; siRNA 4 GAACCUAACUGGACUUUAAUU. For *EZH2* siRNA, the following sequence was used: GAGGUUCAGACGAGCUGAUUU.

shRNA knockdown and western blot analysis. Cells were seeded at 50–60% confluency, incubated overnight, and transfected with EZH2 or nontargeting shRNA lentiviral constructs as described in for 48 h. GFP$^+$ cells were drug-selected using 1 μg/ml puromycin. RNA and protein were harvested for PCR and western blot analysis according to standard protocols. For western blot analysis, PVDF membranes (GE Healthcare) were incubated overnight at 4 °C with either *EZH2* mouse monoclonal (1:1,000, BD Biosciences, no. 612666), or *B-actin* (Abcam, ab8226) for equal loading.

Gene expression profiling. Agilent Whole Human Genome Oligo Microarray was used for cDNA profiling of PCAT-1 siRNA knockdown samples or nontargeting control according to standard protocols_ENREF_50. All samples were run in technical triplicates against nontargeting control siRNA. Expression array data was processed using the SAM method[47] with an FDR ≤ 0.01. Up- and downregulated probes were separated and analyzed using the DAVID bioinformatics platform[48].

ChIP. Assays were done as previously described[25], where 4–7 μg of the following antibodies were used: IgG (Millipore, PP64), SUZ12 (Cell Signaling, no. 3737) and SUZ12 (Abcam, ab12073). ChIP-PCR reactions were done in triplicate with SYBRGreen using 1:150th of the ChIP product per reaction.

***In vitro* translation.** Full-length *PCAT-1,* Halo-tagged *ERG* or *GUS* positive control were cloned into the PCR2.1 entry vector (Invitrogen) and *in vitro* translational assays were done using the TnT Quick Coupled Transcription/Translation System (Promega) with 1 mM methionine and Transcend Biotin-Lysyl-tRNA (Promega) according to the manufacturer's instructions.

Bioinformatic analyses. Sequencing reads were aligned with TopHat[19], and *ab initio* assembly was performed with Cufflinks[3]. Transcriptome libraries were merged and statistical classifiers were developed and employed to filter low-confidence transcripts. Nominated transcripts were compared to UCSC, RefSeq, Vega, Ensembl and ENCODE database, and coding potential was determined with the txCdsPredict program from UCSC. Transcript conservation was determined with the SiPhy package. Differential expression analysis was performed using SAM methodology, and outlier analysis using a modified COPA method. See the **Supplementary Methods** for details on the bioinformatics methods used.

Statistical analyses for experimental studies. All data are presented as means ± s.e.m. All experimental assays were performed in duplicate or triplicate. Statistical analyses shown in figures represent Fisher's exact tests or two-tailed Student *t*-tests, as indicated. For details regarding the statistical methods employed during RNA-Seq and ChIP-Seq data analysis, see **Supplementary Methods**.

46. Maher, C.A. *et al.* Chimeric transcript discovery by paired-end transcriptome sequencing. *Proc. Natl. Acad. Sci. USA* **106**, 12353–12358 (2009).
47. Tusher, V.G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA* **98**, 5116–5121 (2001).
48. Dennis, G. Jr. *et al.* DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol.* **4**, 3 (2003).

doi:10.1038/nbt.1914